Improving CAPTCHA Robustness via Controlled Image Corruptions (Student Abstract)

Suchetan G. Uppur, Ashish Kumar, Akshay Agarwal

Trustworthy BiometraVision Lab, IISER Bhopal, India
{suchetan21, ashish21, akagarwal}@iiserb.ac.in

Abstract

The Completely Automated Public Turing test to Tell Computers and Humans Apart (CAPTCHA) is widely deployed on the web as a security mechanism to distinguish humans from automated bots. However, their robustness is being challenged by the rapid advancements in AI, with models capable of near-human level character recognition rendering CAPTCHA obsolete. This research aims to systematically study the effect of multiple image corruptions, including elastic transformations, blur, noise, and occlusions, on human readability and automated solvers in text-based CAPTCHA recognition. We conduct experiments on multimodal large language models (MLLMs), a traditional deep learning-based optical character recognition (OCR) system, and human subjects. Using an existing CAPTCHA dataset and artificially corrupted versions, we analyze the recognition performance of AI models and humans, identifying vulnerabilities and patterns of robustness. The findings will contribute to a better understanding of CAPTCHA vulnerabilities and explore potential methods to increase the robustness of CAPTCHA in the era of advanced AI models.

Introduction

CAPTCHAs serve as a critical barrier against automated systems in online security. However, with the rapid advancement of multimodal large language models and deep learning-based OCR systems, the effectiveness of CAPTCHA is increasingly uncertain. Recent work, such as MCA-Bench (Wu et al. 2025), has provided a large-scale multimodel benchmark to evaluate robustness across diverse types of CAPTCHA. Similarly, IllusionCAPTCHA (Ding et al. 2025) introduced a "Human-Easy but AI-Hard" paradigm using visual illusions, showing that AI models struggle with such challenges while humans perform better. However, both approaches overlook how existing CAPTCHA behaves under image corruptions. Our work aims to fill this gap by comprehensively evaluating CAPTCHA robustness under controlled scenarios of image corruption. We test the recognition performance of humans and other AI methods. We enhance CAPTCHA robustness against attacks by leveraging controlled image corruptions. Our study highlights current limitations and provides insights into designing the next generation of text-based CAPTCHA.

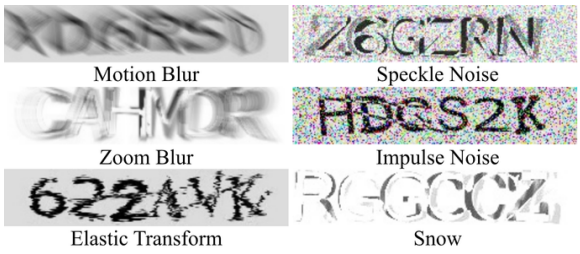

Figure 1: Corrupted versions of CAPTCHA with various distortions like blur, noise, warping, and overlaps to challenge automated recognition. Can you recognise the text in the images? Uncorrupted versions are given at the end!

Proposed Technique for Enhancing CAPTCHA Robustness

To mitigate the vulnerability created from automated attacks by multimodal large language models (MLLMs), we introduce a corruption-based augmentation strategy for CAPTCHA generation to enhance their robustness against machine-based solvers. The proposed approach generates corrupted CAPTCHA images designed to be human-solvable yet AI-resistant. Figure 1 shows a few such corrupted images. We evaluate their effectiveness by measuring solver accuracy and conducting a comparative user study with human participants. Our experiments test the recognition accuracy of CAPTCHA images against three modalities: multimodal large language models (MLLMs), a deep learning-based optical character recognition (OCR) method, and human participants. While our primary focus lies in evaluating MLLM and human performance, we also test a simple OCR model to capture the differences between traditional and modern text recognition methods.

Implementation Details

For our experiments, we utilize the Kaggle CAPTCHA Image Dataset (Bergmann 2025), from which we randomly sample and create an augmented dataset of 120 images.

	Corruption Severity				
Corruption Type	**1**	**2**	**3**	**4**	**5**
Motion Blur	100	75.0	62.5	**25.0**	**25.0**
Elastic Transform	**62.5**	**62.5**	37.5	62.5	50.0
Zoom Blur	87.5	**62.5**	**25.0**	**25.0**	**25.0**
Impulse Noise	67.5	75.0	67.5	67.5	**25.0**
Shot Noise	87.5	100	75.0	37.5	37.5
Snow	87.5	67.5	75.0	50.0	37.5
Speckle Noise	100	87.5	50.0	37.5	37.5

Table 1: Performance (% accuracy) of MLLMs on CAPTCHA recognition under increasing corruption severity, showing performance degradation across different distortion types. **Accuracy on uncorrupted images: 90%**.

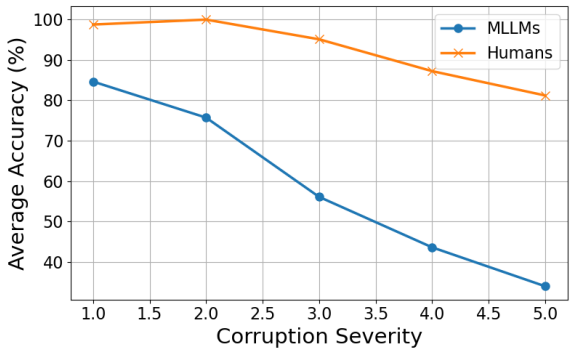

Figure 2: Comparison of MLLM and human performance across corruption severity levels, showing a sharp decline in MLLM accuracy while human accuracy stays high.

Additionally, we generate corrupted versions of these images using the ImageCorruption library (Hendrycks and Dietterich 2019), applying seven different corruptions, elastic transform, snow, zoom blur, impulse noise, motion blur, shot noise, snow, and speckle noise, at five levels of severity, to systematically evaluate the impact of distortions on CAPTCHA recognition. For testing on MLLMs, we use Gemini-2.5 Pro and GPT-5, prompted to recognize a given CAPTCHA image using the prompt: ***"Can you identify the text in the given image?"***. We intentionally use a short and neutral prompt that does not mention CAPTCHA, as MLLM providers often restrict responses to such prompts to discourage the use of their model for fraudulent activities. Furthermore, for testing on OCR, we use the Python Tesseract library (Hoffstaetter 2025), which relies on the open-source LSTM-based Tesseract OCR engine for recognition. For the human study, we included 23 participants aged 20 to 25. Each participant is presented with 20 randomly sampled CAPTCHA images, comprising both uncorrupted and corrupted examples, at five levels of corruption severity.

Experimental Results and Analysis

The results of MLLMs are reported in Table 1, and the accuracies are computed by combining the outcomes of both MLLM models. Table 2 shows human accuracy when recognizing the same CAPTCHA images that MLLMs are tested

	Corruption Severity				
Corruption Type	**1**	**2**	**3**	**4**	**5**
Motion Blur	100	100	90.9	91.7	87.0
Elastic Transform	100	100	91.7	72.7	91.3
Zoom Blur	100	100	**81.8**	**65.2**	65.2
Impulse Noise	100	100	100	100	66.7
Shot Noise	100	100	100	100	**58.3**
Snow	**91.7**	100	100	100	100
Speckle Noise	100	100	100	100	100

Table 2: Human performance (% accuracy) in CAPTCHA recognition across corruption types and severities, showing consistent performance with slight drops at extreme distortion levels. **Accuracy on uncorrupted images: 100%**.

Figure 3: Uncorrupted version of CAPTCHA of Figure 1.

on. The lowest accuracy per severity level is highlighted in bold. We found the performance of Tesseract OCR to be extremely poor (∼0% accuracy) even on uncorrupted images, proving that the design of CAPTCHA (without corruptions) is strong against older OCR-based recognition models. The recognition accuracy trend across corruption severity is captured in Figure 2, which shows that human performance on the corrupted images remains consistently higher than that of MLLMs, indicating that the corruptions are beneficial in making text-based CAPTCHA harder for MLLMs to solve. Moreover, applying similar corruptions to image-based CAPTCHAs may also lead to improved CAPTCHA robustness, as such perturbations disrupt the perceptual embeddings that MLLMs rely on, leading to degraded recognition, reasoning, and confidence (Cui et al. 2024). However, we do not verify this empirically, and it remains to be tested how corruptions affect MLLM performance on CAPTCHA based on other modalities, such as image-based CAPTCHA. Figure 3 showcases the uncorrupted version of CAPTCHA images presented in Figure 1. While Figure 3 demonstrates the ease of readability due to high quality, the understanding of corrupted CAPTCHA by humans highlights how we can restrict its solvability by AI experts, including LLMs/VLM/MLLMs.

Conclusion and Future Work

CAPTCHA remains widely used for bot attack prevention, but advanced MLLMs can easily break traditional text-based CAPTCHAs. We demonstrate that image corruptions can strengthen CAPTCHA security while remaining readable to humans. Future work will explore additional corruption techniques, test against a broader range of AI models, conduct larger human usability studies, and design distortions that are **"Human Easy but AI Hard"**.

References

Bergmann, J. 2025. CAPTCHA Image Dataset. https://www.kaggle.com/datasets/johnbergmann/captcha-image-dataset. Online; accessed September 2025.

Cui, X.; Aparcedo, A.; Jang, Y. K.; and Lim, S.-N. 2024. On the robustness of large multimodal models against image adversarial attacks. In *IEEE/CVF CVPR*, 24625–24634.

Ding, Z.; Deng, G.; Liu, Y.; Ding, J.; Chen, J.; Sui, Y.; and Li, Y. 2025. IllusionCAPTCHA: A CAPTCHA based on visual illusion. In *ACM on Web Conference 2025*, 3683–3691.

Hendrycks, D.; and Dietterich, T. 2019. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*.

Hoffstaetter, S. 2025. Python Tesseract. https://pypi.org/project/pytesseract/. Online; accessed September 2025.

Wu, Z.; Xue, Y.; Wei, X.; and Song, Y. 2025. MCA-Bench: A Multimodal Benchmark for Evaluating CAPTCHA Robustness Against VLM-based Attacks. *arXiv preprint arXiv:2506.05982*.